# Bayesian Classification and Non-Bayesian Label Estimation via EM Algorithm to Identify Differentially Expressed Genes: a Comparative Study

Marília Antunes Lisete Sousa

Department of Statistics and Operation Research and CEAUL, University of Lisbon,

**Abstract**

Gene classification problem is studied considering the ratio of gene expression levels, $X$, in two-channel microarrays and a non-observed categorical variable indicating how differentially expressed the gene is: *non differentially expressed*, *down-regulated* or *up-regulated.* Supposing $X$ from a mixture of gamma distributions, two methods are proposed and results are compared. The first method is based on a hierarchical Bayesian model. The conditional predictive probability of a gene to belong to each group is calculated and the gene is assigned to the group for which this conditional probability is higher. The second method uses EM algorithm to estimate the most likely group label for each gene, that is, to assign the gene to the group which contains it with the higher estimated probability.

**Keywords:** Differential expression, EM algorithm, hierarchical Bayesian model, mixture models.

## 1 Introduction

The level of DNA transcripted to RNA in a normal organism and in a mutation may be compared using global techniques for gene expression (genomics) as microarrays. These techniques have an enormous potential for the comprehension of genes' regulatory function, their interactions and pathophysiological mechanisms. Thus, it is possible to evaluate the impact of a particular genetic change in the global genome expression level (Stekel, 2003).

A typical microarray experiment results in thousands genes expression levels to be measured simultaneously, but with few replicates for each gene, leading to an extreme multiple testing situation (Dudoit *et al.*, 2002; Dudoit *et al.*, 2003 ). A possibility is to adopt some assumptions about the distribution of the expression measures such as within-gene log-normality and independence of replicate measurements on a null hypothesis of equivalent expression (Newton *et al.*, 2004). This assumption is very convenient since a lot of theory is available. However, we should notice that the distribution we are interested in from the point of view of inference is not of the gene expressions over the array but the distribution of the single genes across arrays, usually only few being available. The strictly positive nature of gene expression values makes a normal distribution an unlikely candidate. However, experience showed that after a logarithmic transformation, many gene expressions are indistinguishable from normal distribution (Wit and McClure, 2004).

Since the main objective is to identify genes that are differentially expressed (*DE*), many of the existing methods produce a ranking of the genes according to their expression levels based on $p$-values. The $p-$values are calculated based on the null hypothesis that each gene is not *DE* in the two samples. In such cases a suitable cutoff is needed for the detection of *DE* genes. Due to the difficulty in determining this cutoff point, the top genes are chosen by ranking the evidence favoring differential expression, e.g., the top 50, 100 or 150 genes are selected, as it is often necessary to report a short list of *DE* genes (Lönnstedt and Speed, 2002; Newton *et al.*, 2004). As Reiner *et al.* (2003) refer, the probability that a false identification (type I error) is commited can increase sharply when the number of tested genes is large. When dealing with microarray data, it is appropriate to emphasize the proportion of error among the identified differently expressed genes. The expectation of this proportion is the false discovery rate (FDR) of Benjamini and Hochberg (1995). Obviously, it is desirable to keep FDR as low as possible but it is also necessary to keep in mind that the imposition of too high thresholds may lead to type II errors, which are to consider in this context since the identification of *DE* genes has the main purpose of extracting genes which are potential candidates for further investigation. Hence, several of type I erroneous decisions (to consider as *DE* genes which are not) will not distort the conclusions at early stages of the investigation as long as this proportion remains small.

Another aspect to consider is the fact that in almost all experiences involving microarrays, the number of replicates is considerably small, often not more than four. Hence, when dealing with microarray data, not only multiple testing is a problem but also the lack of robustness constitutes an important aspect to consider. In such small samples, outliers have enormous effect on the results since both very large and very small values for the mean can be driven by the presence of outliers in the data, which occurs quite frequently in this context (Lönnstedt and Speed, 2002). Although normalizing the data along the slides helps to reduce discrepancies, this procedure does not reduce substantially the effect of such observations in the results. In this work we propose also to consider the median of the intensity ratios along the slides and use it for classification purposes.

An assumption that became popular in microarray data analysis is the discrete mixture model (Newton *et al.*, 2001; Efron *et al.*, 2001; Allison *et al.*, 2002; Bröet *et al.*, 2002; Lee *et al.*, 2002; Lönnstedt and Speed, 2002; Pan, 2002; Kendziorski *et al.*, 2003; Storey and Tibshirani, 2003; Dean and Raftery, 2005). A wide variety of mixture models have been proposed, each of which can be expected to provide a different classification of the genes in any given data set (Lewin *et al.*, 2007). In such models, mixtures can be specified directly at the data level. Our model fits into this group.

In this work, the Bayesian approach constitutes an alternative to these methods. We propose a simple model at the data level. Based on a mixture of Gamma distributions, the classification method described here associates each gene to the "expression category" with higher predictive conditional probability. Considering a similar model structure, but from a frequentist point of view, we use EM algorithm to estimate the most likely group label for each gene, that is, to assign the gene to the group which contains it with the higher estimated probability.

Both methods are applied medians and averages calculated using the HIV dataset (two-channel microarrays) available in library `NUDGE` in `R`, and results are compared. A comparison with other methods is also provided.

## 2 Statistical methods

Two channel microarray data consists in intensities, say $R$ and $G$, of each fluorophore (Cy5 and Cy3 dies respectively) used to label the samples to be compared. Consequently, raw data arising from microarrays is strictly positive. Furthermore, in each experiment, thousands of genes are spotted in the slides but few slides (and hence replicates) are done. For these reasons, the problem of identifying *DE* genes is not only a multiple testing problem, but also a small sample one. Early analyses of microarray data relied first on the use of the t-test and later on various transformations of the t-test (Lewin *et al.*, 2007 and references therein), and therefore normality of the data is a prerequisite when adopting these approaches. To fulfill this condition, log-intensities ($\log R$ and $\log G$) or $\log \frac{R}{G}$ have been used.

Other approaches focused on the ratio $\frac{R}{G}$ and used bayesian hierarchical models model $\rho = \frac{\mu_r}{\mu_g}$, the ratio of the mean values of the red and green intensities at a given spot (Newton *et al.*, 2001; Kendziorski *et al.*, 2003; Newton and Kendziorski, 2003)

In this work we chose to use raw data and to consider as measure of the genes' expression, $X$, the median of the ratios $\frac{R}{G}$ measured along the slides for each gene and to consider $X$ from a gamma distribution, with different parameters for genes *down-regulated*, *non-DE* and *up-regulated*. The reasons behind our choice are the following:

- since the number of replicates is very small and because in this kind of data outliers are frequent, to consider medians instead of averages constitutes a way of avoiding the effect of possible outliers. In very small samples, e.g. in a sample of size four, the median will use almost the same amount of information as the average, with the advantage of giving more importance to the main part of the data by being resistant to the effect of the outliers;

- gamma is a very flexible model for strictly positive data, accomodating many possibilities to combine location and dispersion. Also, it's skewness allows to accomodate well higher values, which are usual for the *up-regulated* genes.

## 2.1 Bayesian Classifier

The Bayesian classifier we propose can be used in a wide variety of scenarios (Antunes *et al.*, 2006; Fonseca *et al.*, 2007) and is suitable for the purpose of identifying *DE* genes. It is built assuming that there is a continuous characteristic, $X$, which is measurable for every element in the population in study. The population is considered to be divided in $J$ groups and $X$ is assumed to behave in a different manner, depending on the group the individual belongs to, and therefore it is useful for the purpose of classification. In the present context, $X$ is a measure of the genes' expression level and $J = 3$, the groups corresponding to: *down-regulated*, *non-DE* and *up-regulated*.

The method starts from probabilities calculated based on prior assumptions and then they are updated based on the observed data. The classification rule is based on the predictive conditional distributions. This idea fits in a supervised classification context (Dudoit and Fridlyand, 2003) and the ideal situation would be to have a training set (a distinct from our data but alike group of genes already known to be *down-regulated*, *non-DE* or *up-regulated*), based on which the classification rule would be built and only afterwards applied to the data. Since this generally does not happen, the data is pre-classified based on general information and beliefs and used to build the classification rule.

Consider $\mathcal{D} = \{(x_1, t_1), \ldots, (x_n, t_n)\}$ the available data, where $(x_i, t_i)$, corresponds to the data for the $i - th$ gene and $n$ is the number of genes in the study. Variable $T$ is a categorical variable, assuming values (labels) from 1 to $J = 3$, indicating to which group the gene belongs to.

We propose a Bayesian hierachical model at the data level. This modelling structure is lighter than the generally presented by authors using Bayesian methods (e.g. Newton *et al.* 2001; Lönnstedt and Speed, 2002; Lewin *et al.*, 2007). Since we are considering a strictly positive variable for the gene expression intensity, the following distributions are considered:

- $T \sim$ dcat$(\pi_1, \pi_2, \pi_3)$, that is, $T$ follows a categorical discrete distribution

$$T : \left\{ \begin{array}{ll} j, & j = 1, 2, 3 \\ \pi_j & \end{array} \right. ,$$

where $\pi_j = P[T = j]$ and $\sum_{j=1}^{3} \pi_j = 1$;

- Conditional on $T = j$, $X$ follows a Gamma distribution with shape parameter $\alpha_j$ and scale parameter $\beta_j$, $X_{|T=j} \sim Gamma(\alpha_j, \beta_j)$, $j = 1, 2, 3$, that is,

$$p(x|T = j, \alpha_j, \beta_j) = \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} x^{\alpha_j - 1} e^{-\beta_j x}, \quad x > 0, \alpha_j > 0, \beta_j > 0;$$

- $(\pi_1, \pi_2) \sim Dirichlet(a_1, a_2, a_3)$;
- $\alpha_j, j = 1, 2, 3$, are considered known for mathematical and computational simplicity as is allows to find a closed form for the predictive distributions, which would not be possible otherwise;
- $\beta_j \sim Gamma(g, h)$, $j = 1, 2, 3$, all independent and independent from $\pi_1$ and $\pi_2$;
- $g$, $h$, $a_1$, $a_2$ and $a_3$ are the hyperparameters of the model.

The predictive distribution of $X$ given the value of $T$, $p(x|\mathcal{D}, T = j)$, can be seen as a way to estimate the distribution of $X$ within each group $j$. Plotting these three curves together allow us to compare the groups, as they are useful for descriptive purposes. Namely, they give a clear picture of the way the values of $X$ distribute within and along the groups. The points where these curves intersect define the limits of the regions where elements from different groups coexist. Conversely, the predictive conditional distribution of $T$ given $X = x$ calculated for each group $j$, $p(T = j|\mathcal{D}, x)$, is a function of $x$ and shows how the probability of a gene to belong to group $j$ evolves as a function of the gene's value for the variable $X$. Hence, it can be used to determine to which group the gene is more likely to belong to given $x$, the observed value of $X$. Also, drawing these curves should permit to identify intervals for $x$ over which each of the groups is the most likely to contain the gene. The intersection points can be taken as cutoff points, determining such regions in terms of values of $X$.

### The predictive conditional distribution of T

The predictive conditional distribution of $T$ given $x$ is given by:

$$P(T = j|\mathcal{D}, x) = \frac{p(x|\mathcal{D}, T = j) P(T = j|\mathcal{D})}{p(x|\mathcal{D})} \tag{1}$$

Posterior distributions of $(\pi_1, \pi_2)$ and $\beta_j$ are known to be Dirichlet and Gamma, that is,

$$(\pi_1, \pi_2)|\mathcal{D} \sim Dirichlet(n_1 + a_1, n_2 + a_2, n_3 + a_3)$$

and

$$\beta_j|\mathcal{D} \sim Gamma(G_j, H_j),$$

where $G_j = n_j \alpha_j + g$ and $H_j = \sum_{k=1}^{n_j} x_{jk} + h$, for $j = 1, 2, 3$, where $x_{jk}$ represents the $k-$th observation in group $j$, $n_j$ representing the number of genes pre-classified in group $j$. A posteriori, the $\beta_j$ are all independent and independent of $\pi_i$ for all $i = 1, 2, 3$ and $j = 1, 2, 3$.

The marginal predictive distribution of $X$ is obtained calculating

$$p(x|\mathcal{D}) = \sum_{j=1}^{3} p(x|\mathcal{D}, T = j) p(T = j|\mathcal{D}) \tag{2}$$

where $p(x|\mathcal{D}, T = j)$ is the predictive conditional distribution of $X$ given $T = j$, and $p(T = j|\mathcal{D})$ is the predictive distribution of $T$. These are respectively given by

$$
\begin{aligned}
p(x|\mathcal{D}, T = j) &= \int_0^\infty p(x|T = t_j, \alpha_j, \beta_j) p(\beta_j|\mathcal{D}) d\beta_j \\
&= \frac{\Gamma(\alpha_j + G_j)}{\Gamma(\alpha_j)\Gamma(G_j)} \times \frac{H_j^{G_j} x^{\alpha_j - 1}}{(x + H_j)^{G_j + \alpha_j}}, \quad 0 < x < \infty,
\end{aligned} \tag{3}
$$

and

$$
\begin{aligned}
P(T = j|\mathcal{D}) &= \int_{\mathcal{P}} P(T = j) p(\pi_1, \pi_2|\mathcal{D}) d\pi_1 d\pi_2 d\pi_3 \\
&= \frac{n_j + a_j}{n + a_1 + a_2 + a_3}
\end{aligned} \tag{4}
$$

where $\mathcal{P} = [0, 1] \times [0, 1] \times [0, 1]$.

Now, the predictive conditional distribution of $T$ given $x$ can be written:

$$
P(T = j|\mathcal{D}, x) = \frac{\frac{\Gamma(\alpha_j + G_j)}{\Gamma(\alpha_j)\Gamma(G_j)} \times \frac{H_j^{G_j} x^{\alpha_j - 1}}{(x + H_j)^{G_j + \alpha_j}} \times \frac{n_j + a_j}{n + a_1 + a_2 + a_3}}{\sum_{l=1}^{3} \frac{\Gamma(\alpha_l + G_l)}{\Gamma(\alpha_l)\Gamma(G_l)} \times \frac{H_l^{G_l} x^{\alpha_l - 1}}{(x + H_l)^{G_l + \alpha_l}} \times \frac{n_l + a_l}{n + a_1 + a_2 + a_3}} \tag{5}
$$

### The classification rule

As said above, the predictive conditional probability functions $p(T = j|\mathcal{D}, x)$ show how the probability of a gene for which $X = x$ to belong to group $j$ evolves as a function of $x$. Therefore these functions can be used for the purpose of classification, with a gene for which $X = x$ being classified in group $j$ if

$$
j = \underset{j}{\text{argmax}}\{P(T = j|\mathcal{D}, x), j = 1, 2, 3\}.
$$

Plotting the functions $p(T = j|\mathcal{D}, x), j = 1, 2, 3$ together allows us to have a clear picture of the way the "allocation" probabilities compete given $x$. If $X$ is such that it shows typically increasing values as the values for the labels increase, it is expected the data to permit to find values $a$ and $b$, $0 < a < b$, such that:

- for $x \in [0, a]$
$$
P[T = 1|\mathcal{D}, x] > P[T = 2|\mathcal{D}, x] > P[T = 3|\mathcal{D}, x]
$$
- for $x \in (a, b)$
$$
P[T = 2|\mathcal{D}, x] > P[T = 1|\mathcal{D}, x] \text{ and } P[T = 2|\mathcal{D}, x] > P[T = 3|\mathcal{D}, x]
$$
- for $x \in [b, +\infty)$
$$
P[T = 3|\mathcal{D}, x] > P[T = 2|\mathcal{D}, x] > P[T = 1|\mathcal{D}, x],
$$

where $a$ and $b$ are the solutions of the equations $P[T = 1|\mathcal{D}, x] = P[T = 2|\mathcal{D}, x]$ and $P[T = 2|\mathcal{D}, x] = P[T = 3|\mathcal{D}, x]$, respectively. This situation is clearly illustrated in the two graphs at the top of Fig. 2.

In conclusion, given gene for which $X = x$, the classification rule is

$$
\begin{array}{ll}
\text{classify in group 1,} & \text{if } x \in [0, a]; \\
\text{classify in group 2,} & \text{if } x \in (a, b); \\
\text{classify in group 3,} & \text{if } x \in [b, +\infty).
\end{array} \tag{6}
$$

## 2.2 Non-Bayesian classifier - mixture model and label estimation via EM algorithm

In the presence of a dataset assumed to contain data from different groups, it is common to consider a mixture model. Being so, each data point has membership in one of the distributions. If the classification of the data in the groups is unknown, the problem of estimating the model parameters can be seen as a missing data problem. EM-algorithm is an iterative way of computing the missing memberships of data points in the chosen distributions and estimating the model parameters. Now we consider that the only available data is $\{x_1, \dots, x_n\}$.

Using the same dataset, Dean and Raftery (2005) consider the existence of two groups (*DE* and *non-DE* genes) and use a Normal-Uniform mixture model to model the data. The data is log-transformed and EM-algorithm is performed to detect differential gene expression.

Our model is a mixture of three gamma models. We model the gene expression as existing three different groups - *down-regulated*, *non-DE* and *up-regulated* - each group being modeled by its own density. The data as a whole is then modeled as a weighted mixture of these three densities, where the weights correspond to the prior probabilities of being in each of the three groups. Here, the idea is to estimate the posterior probabilities of a gene to belong to each group and classify it in the group presenting the higher probability. The procedure is applied to all the genes.

Consider the groups, $G_k$, $k = 1, 2, 3$. Let

$$\pi_k = P(X_i \in G_k), \quad i = 1, \dots, n, \tag{7}$$

where $\pi_k \in (0, 1)$ for $k = 1, 2, 3$, $\sum_{k=1}^{3} \pi_k = 1$ and $X_i$ represents the expression level for the gene $i$. Conditional on the group, the gene expression level follows a gamma distribution, $X|_{X \in G_k} \sim Gamma(\alpha_k, \beta_k)$. Hence,

$$f(x) = \sum_{k=1}^{3} \pi_k f(x|\alpha_k, \beta_k) \tag{8}$$

where

$$f(x|\alpha_k, \beta_k) = \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} x^{\alpha_k - 1} e^{-\beta_k x}, \quad x > 0, \alpha_k > 0, \beta_k > 0. \tag{9}$$

EM algorithm is an iterative procedure in two steps: the Expectation step and the Maximization step. At each iteration, new membership values are computed in the E-step and, using this values, new estimates for the model parameters are computed in the M-step. With the new values for the model parameters, we proceed back to the E-step to recompute new membership values. The procedure is repeated until there is no substantial change in the mixture model parameters. Bellow we describe the $j$-th iteration of the algorithm.

- Expectation step:

  For each data point $x_i$, $i = 1, \dots, n$, the membership value for the group $k$, $k = 1, 2, 3$, is given by

  $$\hat{y}_{k,i}^{(j)} = \frac{\hat{\pi}_k^{(j-1)} f(x_i|\hat{\alpha}_k^{(j-1)}, \hat{\beta}_k^{(j-1)})}{\sum_{l=1}^{3} \hat{\pi}_l^{(j-1)} f(x_i|\hat{\alpha}_l^{(j-1)}, \hat{\beta}_l^{(j-1)})}. \tag{10}$$

  The membership value $y_{k,i}$ can be seen as the "contribution" of group $k$ to the density function of the mixture model calculated for the data point $x_i$.

- Maximization step:

  With the expectation values in hand for the group membership, we recompute plug-in estimates for the distribution parameters. The $\hat{\pi}_k^{(j)}$ are given by the average of the membership values calculated in the E-step and $\hat{\alpha}_k^{(j)}$ and $\hat{\beta}_k^{(j)}$ are calculated using the method of moments. Therefore, it is necessary to start by recomputing the two first empirical moments for each group, $m_{1,k}^{(j)}$ and $m_{2,k}^{(j)}$ respectively.

$$\hat{\pi}_k^{(j)} = \frac{\sum_{i=1}^n \hat{y}_{k,i}^{(j)}}{n}, \quad k = 1, 2, 3. \tag{11}$$

$$m_{1,k}^{(j)} = \frac{\sum_{i=1}^n \hat{y}_{k,i}^{(j)} x_i}{\sum_{i=1}^n \hat{y}_{k,i}^{(j)}}, \quad m_{2,k}^{(j)} = \frac{\sum_{i=1}^n \hat{y}_{k,i}^{(j)} x_i^2}{\sum_{i=1}^n \hat{y}_{k,i}^{(j)}}, \quad k = 1, 2, 3. \tag{12}$$

$$\hat{\alpha}_k^{(j)} = \frac{(m_{1,k}^{(j)})^2}{m_{2,k}^{(j)} - (m_{1,k}^{(j)})^2}, \quad \hat{\beta}_k^{(j)} = \frac{m_{1,k}^{(j)}}{m_{2,k}^{(j)} - (m_{1,k}^{(j)})^2} \quad k = 1, 2, 3. \tag{13}$$

To start the algorithm, initial values for the membership variables are needed. We chose them the following way:

$$\hat{y}_{1,i}^{(0)} = \begin{cases} 1, & \text{if } x_i \leq Q_p \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

$$\hat{y}_{2,i}^{(0)} = \begin{cases} 1, & \text{if } Q_p < x_i \leq Q_{1-p} \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

$$\hat{y}_{3,i}^{(0)} = \begin{cases} 1, & \text{if } x_i > Q_{1-p} \\ 0, & \text{otherwise,} \end{cases} \tag{16}$$

with $p$ small, $p \in (0, 0.5)$, where $Q_p$ represents the quantile of order $p$. In principle, $p$ should be chosen quite small since $2p$ will correspond to the proportion of genes accepted a priori as differentially expressed (half of them being *down-regulated* and the other half being *up-regulated*). Using such initial values for the membership variables, the initial values for the mixture model parameters are computed using equations (11), (12) and (13), taking $j = 0$.

As said above, E-step and M-step are repeated until no significant changes occur in the estimates of the mixture model parameters, this is, the iterative procedure stops after iteration $j$ if

$$|\hat{\pi}_k^{(j)} - \hat{\pi}_k^{(j-1)}| < \delta, \quad |\hat{\alpha}_k^{(j)} - \hat{\alpha}_k^{(j-1)}| < \delta \text{ and } |\hat{\beta}_k^{(j)} - \hat{\beta}_k^{(j-1)}| < \delta, \tag{17}$$

for every $k = 1, 2, 3$, and a sufficiently small $\delta$. Otherwise, perform iteration $(j + 1)$.

# 3 Application

## 3.1 The HIV data

The HIV[a] data analyzed here (van't Wout *et al.*, 2003) consists of 4 replicate experiments in order to compare cDNA from CD4+ T cell lines at 1 hour after infection with HIV-1BRU with non-infected cell lines on each slide.

[a] Data available at: `http://expression.microslu.washington.edu/expression/vantwoutjvi2002.html`; R package `NUDGE`.

This dataset contains **13** genes known to be *DE* (positive controls) and **29** genes *non-DE* (negative controls), which are useful to evaluate the performance of the classification procedures, namely to study the sensibility and sensitivity. There are **4608** genes and the 4 replicates have balanced dye swaps. As a measure of comparison between both samples, we consider the ratio between them since Gamma distribution requires positive observations:

$$r_{ij} = ratio = \frac{\text{infected cells expression level (gene } i\text{, replicate } j\text{)}}{\textbf{non} \text{ infected cells expression level (gene } i\text{, replicate } j\text{)}}\,; i = 1, \ldots, 4608, j = 1, \ldots, 4.$$

Preliminary analysis of the data showed that for some genes there were enormous differences in the values of the ratios from slide to slide. Particularly, this happened with one of the negative control genes. Because highly discrepant values cause very significant changes in the value of the mean, means and medians of the ratios were calculated. The classification procedures were performed considering both measures and results were compared. To illustrate these discrepancies, let $a_{ij}$ be the mean intensity of gene $i$ in replicate $j$ over the two samples (infected and non-infected):

$$a_{ij} = \textit{mean intensity} = \frac{\text{infected cells expression level } (i,j) + \textbf{non} \text{ infected cells expression level } (i,j)}{2}.$$

In Fig.1, $m_i = average(r_{ij})$ and $x_i = median(r_{ij})$ are plotted against $c_i = average(a_{ij})$ and $d_i = median(a_{ij})$, respectively. When considering the average as the gene expression value, one of the negative controls is found among the high values of the positive controls. If the median is considered, the two groups (positive and negative controls) separate clearly although the smallest positive control and the highest negative control exhibit very close values.

## 3.2 Results

Both classification procedures were performed using medians ($x_i$) and averages ($m_i$) as input data. In both methods pre-classification of the data was done considering values of $p$ from $0.01$ down to $0.0025$ since it is a general belief that only a very small proportion of the genes show trully differential expression. Results are presented bellow and comparisons are made both between methods and datasets, taking into account the number of genes identified as differentially expressed, the number of false negatives and the number of false positives. In Fig.2, plots of the predictive conditional probabilities for medians and averages can be compared for $p = 0.01$ and $p = 0.0025$ and results are summarized in Table 1. Furthermore, the best preforming method is compared to other methods, namely, NUDGE and EBarrays (Table 2). Because the dataset does not contain genes known to be *down-regulated*, it is not possible to compare the performance of the methods when using each input data nor the proportion of genes pre-classified as *down-regulated*.

### Bayesian Classifier

The application of the Bayesian classifier requires pre-classifying the values in the tails with weight $p$ as *DE* genes:

- *down-regulate*d: $p \times 100\%$;
- *non-DE*: $(1 - 2p) \times 100\%$;
- *up-regulated*: $p \times 100\%$.

For each group, the shape parameter $\alpha_j$ does not enter as a parameter in the hierarchical model and it is considered as a known constant. Because these values are not indeed known, in each group $\alpha_j$ is estimated using the method of moments. The hyperparameters of the model, $g$, $h$, $a_1$, $a_2$ and $a_3$ were considered null in the calculation of the predictive distributions.

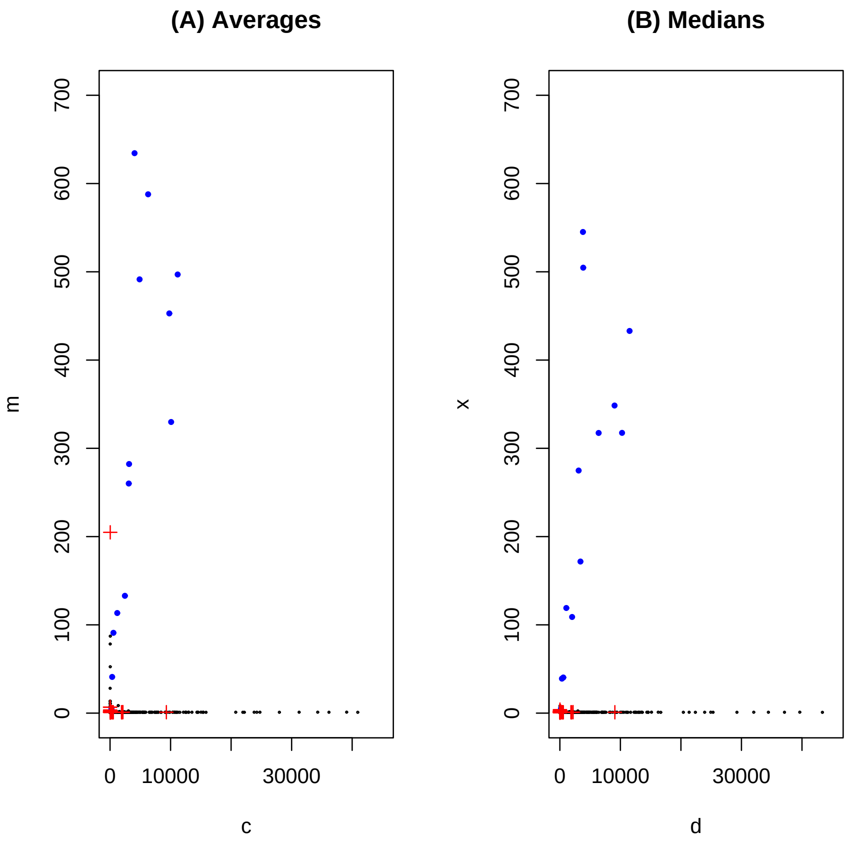

**Figure 1:** Data plots considering averages (A) and medians (B). Averages and medians of the ratios are plotted against averages and medians of mean intensities considering infected and non-infected cells, respectively. Large dots: positive controls; +: negative controls; small dots: non classified genes.

*Input data: medians* $(x_i)$

For every value of $p$ considered, the 13 positive controls were correctly identified. Concerning the negative controls, the best result was obtained for $p = 0.0025$ with only one false positive among the 28 genes selected to be *up-regulated*. The number of genes selected as being *DE* goes from 31(=28 *up-regulated* + 3 *down-regulated*=0.67% of the total number of genes for $p = 0.0025$) to 101(=73 *up-regulated* + 28 *down-regulated*=2.19% of the total number of genes for $p = 0.01$). For $p = 0.0025$ gene $i$ is classified as *down-regulated* if $x_i < 0.33$ and, for $p = 0.01$ gene $i$ is classified as *down-regulated* if $x_i < 0.53$, being classified as *up-regulated* if $x_i > 2.065$ and $x_i > 3.8$ for $p = 0.01$ and $p = 0.0025$, respectively.

*Input data: averages* $(m_i)$

When considering values of $p \geq 0.005$, all the 13 positive controls are correctly identified. Four and three of the negative control genes are classified as *up-regulated* for $p = 0.01$ and $p = 0.005$, respectively. For $p = 0.004$ the number of false positives remains the same as for $p = 0.005$ but there is one false negative as one of the positive control genes is wrongly classified as *non-DE*. For $p \leq 0.003$ the number of false positives drops to one while the number of false negatives remains equal to one. When using the averages as input data for the Bayesian classifier, the number of genes selected as being *DE* goes from 16 ($p = 0.0025$) to 77 ($p = 0.01$), all classified as *up-regulated*, corresponding to 0.35% and 1.67% of the total number of genes, respectively. For $p = 0.0025$ and $p = 0.01$, gene $i$ is classified as *up-regulated* if $x_i > 38.2$ and $x_i > 2.84$, respectively.

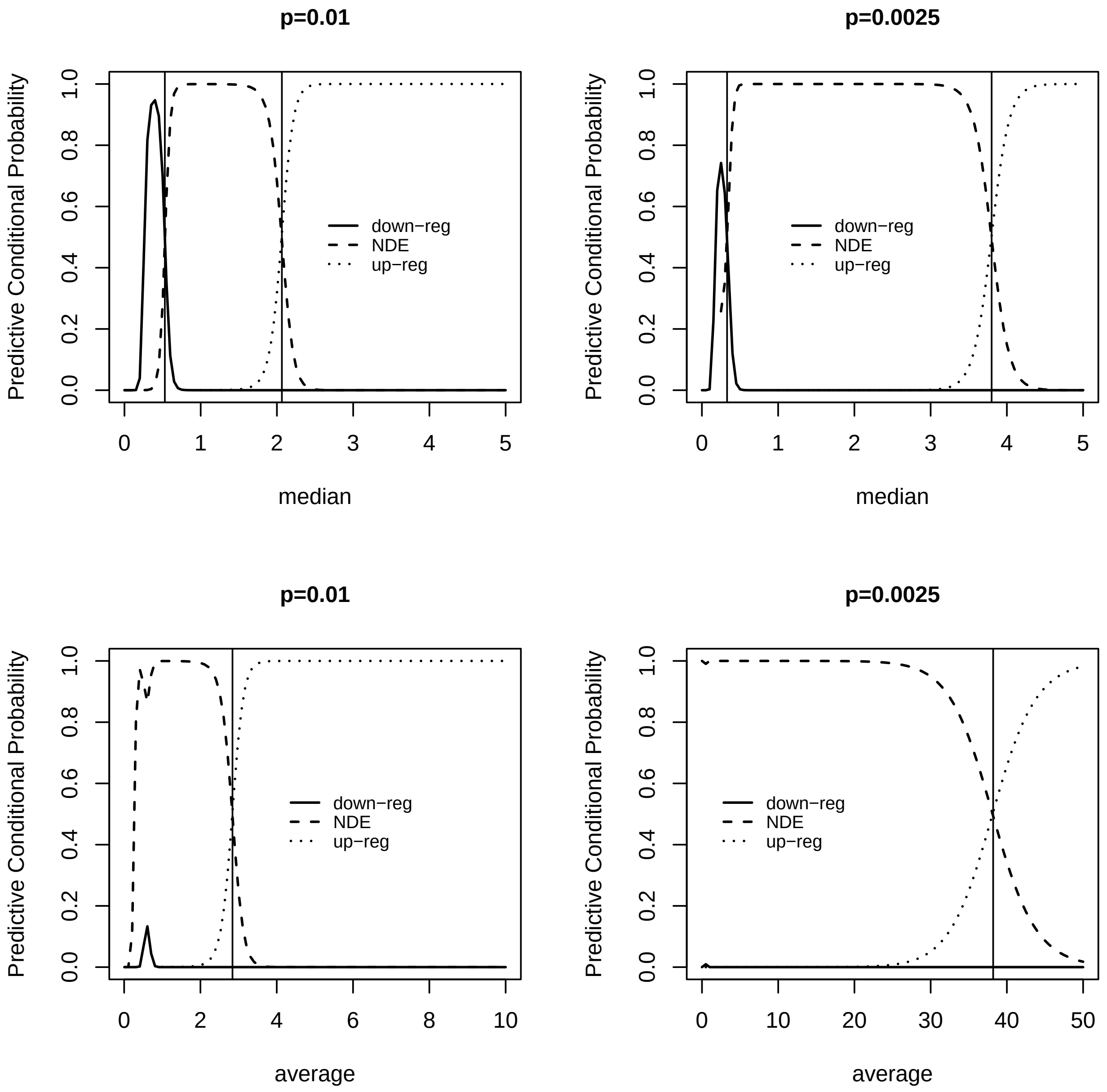

**Figure 2:** Predictive conditional distribution functions.

### Non-Bayesian classifier - mixture model and label estimation via EM Algorithm

EM algorithm was applied considering the same values of $p$ as for the Bayesian classifier and several values of $\delta$ from $10^{-2}$ to $10^{-4}$. For every combination of $p$ and $\delta$, the result was the same. When considering the medians as input data, the method classifies almost two times more genes as *up-regulated* than as *down-regulated* (153 and 82 genes, respectively). When considering the averages as input data, this difference increases immensely: 201 genes classified as *up-regulated* whereas only 13 genes are classified as *down-regulated*. Apart from this aspect, the results were very similar, with a slightly better outcome when using medians as input data concerning the false discovery rate.

*Input data: medians* $(x_i)$

All the 13 positive controls were identified as being *up-regulated* and hence there are no false negatives whereas there are 7 false positives, corresponding to negative controls that were wrongly classified as *up-regulated*. Out of the 4608 genes, 82 were classified as *down-regulated* corresponding to $1.78\%$ of the total number of genes. The number of genes classified as *up-regulated* correspond to $3.32\%$ of the total number of genes.

*Input data: averages* $(m_i)$

Again all the 13 positive controls were identified as being *up-regulated*, and hence there are no false negatives. The number of false positives rises to 8, corresponding to negative controls that were wrongly classified as *up-regulated*. Out of the 4608 genes, only 13 were classified as *down-regulated* corresponding to $0.28\%$ of the total number of genes, whereas $4.63\%$ of the total number of genes is classified as *up-regulated*.

### Comparison between presented classifiers and other methods

Concerning the number of false positives and the total number of genes classified as *DE*, the Bayesian method performs better than using EM algorithm in a non-Bayesian framework. This is valid for both values of $p$, being more evident for $p = 0.0025$. Still, the use of medians $(m_i)$ produces a slight improvement in the number of false positives, except for the Bayesian method, with $p = 0.0025$, where this evidence is stronger due to the false negative obtained for the averages.

| | input data: $x_i$; $p = 0.01$ | | | input data: $x_i$; $p = 0.0025$ | | |
|---|---|---|---|---|---|---|
| Approach | Class. *DE* (*up-reg*) | nr. false neg | nr. false pos | Class. *DE* (*up-reg*) | nr. false neg | nr. false pos |
| Bayesian | 101 (73) | 0 | 3 | 31 (28) | 0 | 1 |
| non-Bayesian (EM) | 235 (153) | 0 | 7 | 235 (153) | 0 | 7 |
| | input data: $m_i$; $p = 0.01$ | | | input data: $m_i$; $p = 0.0025$ | | |
| Approach | Class. *DE* (*up-reg*) | nr. false neg | nr. false pos | Class. *DE* (*up-reg*) | nr. false neg | nr. false pos |
| Bayesian | 77 (77) | 0 | 4 | 16 (16) | 1 | 1 |
| non-Bayesian (EM) | 214 (201) | 0 | 8 | 214 (201) | 0 | 8 |

**Table 1:** Summary of the results obtained from the Bayesian method and the non-Bayesian (EM algorithm). In Class. *DE* column, values in brackets represent the number of genes classified as *up-regulated* among the Class. *DE*.

The idea of using mixture models is given by other authors, as it seem to be a natural and intuitive approach. Two of these approaches are NUDGE and EBarrays. For both, the data are assumed to be generated by a two component mixture model, one component for differentially expressed and the other for non-differentially expressed genes, each with their own distribution. EBarrays has been used in the context of a Bayesian analysis (Newton and Kendziorski, 2003), assuming that the observed ratios had a gamma distribution and its scale parameter itself had a gamma distribution (GG), or, as an alternative assumption, that the observed log ratios were normally distributed and the prior for the mean was normal also (LNN). The posterior probability was then used to make inference about differential expression (Newton *et al.*, 2001). NUDGE uses a simple univariate normal-uniform mixture model, trough the log ratio observations; a normal component for those genes that are not differentially expressed and a uniform component for those that are. The model is estimated by maximum likelihood using the EM algorithm (Dean and Raftery, 2005). For comparisons to NUDGE and EBarrays the best performing method - Bayesian classifier - will be considered.

Table 2 shows the results of all methods for the HIV data. NUDGE and EBarrays Gamma-Gamma had a perfect result for the control genes, with no false positives and no false negatives. As for the Bayesian-Medians model, the EBarrays Lognormal-Normal model had one false positive. In what concerns the total number of genes identified as *DE*, the Bayesian-Averages method equals NUDGE's performace, however is penalized with one false negative and one false positive.

The Bayesian method works better when considering the medians. Even though, one false positive is identified and the total number of *DE* genes is slightly higher then for the other methods. However, it is important to emphasize the fact of all the 13 positive controls being detected, thus it is possible

| Method | Nr False Positives | Nr False Negatives | Class. *DE* |
|---|---|---|---|
| Bayesian - Medians | 1 | 0 | 31 |
| Bayesian - Averages | 1 | 1 | 16 |
| Nudge | 0 | 0 | 16 |
| EBarrays GG | 0 | 0 | 24 |
| EBarrays LNN | 1 | 0 | 19 |

**Table 2:** Number of false positives, false negatives and genes classified as differentially expressed for each methods.

to produce a ranked list of 31 genes for further analysis which includes all the *DE* genes. Besides, the Bayesian method offers some advantages as simplicity, speed and few assumptions.

# 4 Discussion and Outlook

Bayesian classifier revealed to be, for these data, the best method in identifying *DE* genes. Even when starting with $p = 0.01$ for each tail, using the $x_i$ as input data, this method selects 101 *DE* genes (against 235 from the non-Bayesian classifier) with 3 false positives (against 7 from the non-Bayesian classifier). Results for the $m_i$ data are 77, 214, 4 and 8, respectively.

Definitely, the Bayesian method starting with tails weighing $0.25\%$ and using $x_i$ as input data, is the best solution as there are no false negatives, only one false positive and a reasonable number of genes classified as *DE*. Moreover, note that this false positive is the negative control gene which intensities are very high, as observed in Fig.1. Although the summarization through medians was able to separate this gene, it differs less than one unit from the positive control gene showing the smallest value and none of the methods applied was able to identify it correctly.

Considering the importance of selecting only a reasonable number of genes (closer to the probable number of differentially expressed genes), the Bayesian approach seems to be more promising than the the non-Bayesian approach.

The disadvantage of the Bayesian method, compared to the non-Bayesian classifier, is the fact that the classification rule depends on finding a numerical solution for the intersection of the predictive conditional probability functions (5).

An advantage of both approaches is the use of raw data, where no log-transformation nor normalization procedures are applied. In fact, the normalization between arrays was performed but no changes in the results were found.

### Acknowledgements

Research partially funded by the projects FCT/POCI 2010 and PTDC/MAT/64353/2006.

# References

Allison, D.B., Gadbury, G.L., Heo, M., Fernandez, J.R., Lee, C.K., Prolla, T.A. and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* **35**, 1–20.

Antunes M, Andreozzi V, Amaral Turkman MA. (2006). A Note on the Use of Bayesian Hierarchical Models for Supervised Classification. Research Report 13, *Notas e Comunicações do Centro de Estatística e Aplicações da Universidade de Lisboa*. 2006.

Benjamini, Y. and Hochberg, Y.(1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Ser. B* **57**, 289–300.

Bröet, P., Richardson, S. and Radvanyi, F. (2002). Bayesian Hierarchical Model for Identifying Changes in Gene Expression from Microarray Experiments. *Journal of Computational Biology* **9**, 671–683.

Dean, N. and Raftery, A.E. (2005). Normal uniform mixture differential gene expression detection for cDNA microarrays. *BMC Bioinformatics* **6**, 1–14.

Dudoit, S., Yang, Y.H., Callow, M. and Speed, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–139.

Dudoit, S., Shaffer, J.P. and Boldrick, J.C. (2003). Multiple hipothesis Testing in Microarray Experiments. *Statistical Science* **18**, 1, 71–103.

Dudoit, S. and Fridlyand, J. (2003). Classification in microarrays experiments. *Statistical Analysis of Gene Expression Microarray Data. Ed.* Terry Speed. Chapman & Hall.

Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association* **96**, 456,1151–1160.

Fonseca, J.E.,Cavaleiro, J., Teles, J., Sousa, E., Andreozzi, V.L., Antunes, M., Amaral-Turkman, M.A., Canhão H., Mourão, A.F., Lopes, J., Caetano-Lopes, J., Weinmann, P., Sobral M., Nero, P., Saavedra, M.J., Malcata, A., Cruz, M., Melo, R., Braña, A., Miranda, L., V-Patto, J., Barcelos, A., Canas da Silva, J., Santos, L.M., Figueiredo, G., Rodrigues, M., Jesus, H., Quintal, A., Carvalho, T., Pereira da Silva, J.A., Branco, J. and Queiroz, M.V. (2007). Contribution for new genetic markers of rheumatoid arthritis activity and severity: sequencing of the tumor necrosis factor-alpha gene promoter. *Arthritis Research & Therapy* 9:R37 (doi:10.1186/ar2173).

Kendziorski, C.M., Newton, M.A., Lan, H., Gould, M.N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**, 3899–3914.

Lee, M.L.T., Lu, W., Whitmore, G.A. and Beier, D. (2002). Models for microarray gene expression data. *Journal of Biopharmaceutical Statistics* **12**, 1–19.

Lewin, A., Bochkina, N. and Richardson, S. (2007). Fully Bayesian Mixture Model for Differential Gene Expression: Simulations and Model Checks. *Statistical Applications in Genetics and Molecular Biology* **6** (1), article 36.

Lönnstedt, I. and Speed, T. (2002). Replicated Microarray Data. *Statistica Sinica* **12**, 31–46.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. and Tsui (2001). On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *Journal of computational Biology* **8** (1), 37–52.

Newton, M.A. and Kendziorski, C.M. (2003). Parametric Empirical Bayes Methods for Microarrays. In *The analysis of gene expression data: methods and software*. Eds. G. Parmigiani, E.S. Garrett, R. Irizarry and S.L. Zeger, New York: Springer Verlag.

Newton, M., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics* **5**, 155–176.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **18**, 546–554.

Reiner, A., Yekutieli, D. and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 3, 368–375.

Stekel, D. (2003). *Microarray Bioinformatics*. Cambridge University Press.

Storey, J.D. and Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proceedings* *of the National Academy of Sciences* **100**, 9440–9445.

van't Wout, A.B., Lehrma, G.K., Mikheeva, S.A, O'Keeffe, G.C., Katze, M.G., Bumgarner, R.E., Geiss, G.K. and Mullins, J.I. (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4+-T-Cell lines. *The Journal of Virology* **77**, 1392–1402.

Wit, E. and McClure, J. (2004). *Statistics for Microarrays*. Wiley.